NUSC Technical Report 7989
11 June 1987

The Moments of Matched and Mismatched Hidden Markov Models

Roy L. Streit
Submarine Sonar Department

Naval Underwater Systems Center
Newport, Rhode Island / New London, Connecticut

Preface

This report was prepared under NUSC Project No. A75210, "Application of Hidden Markov Models to Transient Classification," Principal Investigator Dr. R. L. Streit. This project is part of the NUSC Independent Research Program sponsored by the Office of Naval Research.

The technical reviewer for this report was Dr. J. P. Ianniello (code 2112).

The author thanks Dr. J. P. Ianniello of the Naval Underwater Systems Center, New London, CT, and Jeff Woodard of Rockwell International Corporation, Anaheim, CA, for their helpful comments and discussions.

Reviewed and Approved: 11 June 1987

W. A. Von Winkle

W. A. Von Winkle
Associate Technical Director for Technology

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) TR 7989	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Underwater Systems Center 6b. OFFICE SYMBOL (If applicable) 214	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) New London Laboratory New London, CT 06320	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61152N PROJECT NO. RR0000-N01 TASK NO. RR0000-N01 WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
THE MOMENTS OF MATCHED AND MISMATCHED HIDDEN MARKOV MODELS

12. PERSONAL AUTHOR(S)
Roy L. Streit

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 1/87 TO 6/87	1987 June 11	40

16. SUPPLEMENTARY NOTATION

17 COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	
12	01		Hidden Markov Models (HMMs); Moments; Log Likelihood Ratio; Receiver Operating Characteristics (ROC); Maximum Likelihood Classification
17	01		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

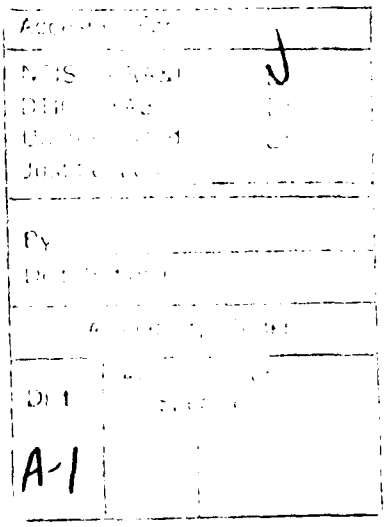
An algorithm for computing the moments of matched and mismatched hidden Markov models from their defining parameters is presented. The ability of the first two moments to adequately describe the probability density function of a maximum posterior likelihood classifier based on hidden Markov models is assessed by examples. These examples include ergodic and nonergodic simulated hidden Markov observations that are matched and mismatched with the posterior likelihood classifier. One example discusses the effect of a noisy discrete communication channel on the posterior likelihood classifier reliability. The examples indicate that the posterior likelihood function is log-normal when the Markov chains are ergodic, and thus the first two moments suffice to describe the required probability density functions. The examples are also of independent interest because they indicate how different internal structures of hidden Markov models impact the performance of a maximum posterior likelihood classifier.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL Roy L. Streit	22b. TELEPHONE (Include Area Code) (203) 440-4906 22c. OFFICE SYMBOL 214

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

TABLE OF CONTENTS



LIST OF ILLUSTRATIONS

LIST OF TABLES

THE MOMENTS OF MATCHED AND MISMATCHED HIDDEN MARKOV MODELS

I. INTRODUCTION

Hidden Markov models (HMMs) are statistical models of nonstationary time series or signals. In speech applications, they are used to characterize the time variation of the short term spectra of spoken words. A specific example is the speaker-independent isolated word recognition (SIIWR) problem, where HMMs characterize the words in a finite-size vocabulary. Different words are characterized by different HMMs.[1]

Every HMM is comprised of two basic parts: a Markov chain and a set of random variables. The Markov chain has a finite number of states, and each state is uniquely associated with one of the random variables. The state sequence generated by the chain is not observable; i.e., the Markov chain is "hidden." At each time $t = 0, 1, 2, \ldots$, the Markov chain is assumed to be in some state; it transitions to another state at time $t + 1$ according to its transition probability matrix. At each time t, one observation is generated by the random variable associated with the state of the Markov chain at time t. The observations are referred to as symbols. If the random variables assume only a finite set of values, the HMM is referred to as a finite symbol HMM. If the random variables assume a continuum of values, the HMM is called a continuous symbol HMM. The full parameter set defining an HMM is comprised of the initial state probability density function of the Markov chain at time $t = 0$, the Markov chain state transition probability matrix, and the probability density functions of each of the random observation variables.

The act of computing specific numerical values for the various parameters of an HMM is called "training," and training is equivalent to solving a mathematical optimization problem to determine maximum likelihood estimates of the HMM parameters.[2] In this paper, it is assumed that the training phase is completed and that the HMMs developed are adequate models for each of the nonstationary time series, or signals, of interest (e.g., the vocabulary words in the SIIWR problem).

During the training phase, a suitable preprocessor is developed to map (or transform) an arbitrary input signal $s(t)$, $t \geq 0$, into a discrete observation sequence $\{O(t),\ t = 1, 2, \ldots\}$. A good description of one way this is done for the SIIWR problem is given in reference 1 (pp. 1077-1078). The preprocessor is also utilized in the classification phase as depicted in figure 1. The observation sequence is truncated to the length T required for the HMMs, where T is a positive integer. The truncated sequence $O_T = \{O(t),\ t = 1, 2, \ldots, T\}$ is then passed to the HMM recognizers. Each HMM recognizer evaluates the posterior likelihood that O_T comes from a time series characterized by that HMM. Denote the i-th hidden Markov model by HMM(i). The i-th recognizer thus computes the posterior likelihood function

$$f_i(O_T) = \Pr[O_T \mid O_T \in \mathrm{HMM}(i)], \quad i = 1, \ldots, p, \tag{1}$$

where $O_T \in \mathrm{HMM}(i)$ denotes the hypothesis that the observation sequence O_T is a realization of HMM(i). The maximum of the p computed posterior likelihoods is assumed to identify, or classify, the original signal $s(t)$. In practice, some kind of tie-breaking rule must be defined and some threshold must be set to identify signals for which HMMs have not been trained. The likelihood function (1) can be computed with only n^2T multiplications (where n is the number of states in the Markov chain) by using the forward-backward algorithm.[2]

The misclassification rate (or false alarm rate) of the system depicted in figure 1 can be estimated by simulation after training is completed. Alternatively, the misclassification rate of signal i as signal j can be determined from the conditional cumulative distribution functions

$$F_{ij}(x) = \Pr[f_i(O_T) < x \mid O_T \in \mathrm{HMM}(j)] \tag{2}$$

by using classical detection and estimation methods to develop receiver-operator characteristics (ROC) curves.[3] The validity of misclassification rates based on $F_{ij}(x)$ depends on the validity of the assumption that the HMMs developed are adequate models for the signals of interest. Agreement between theory and simulation would support the hypothesis that the HMMs

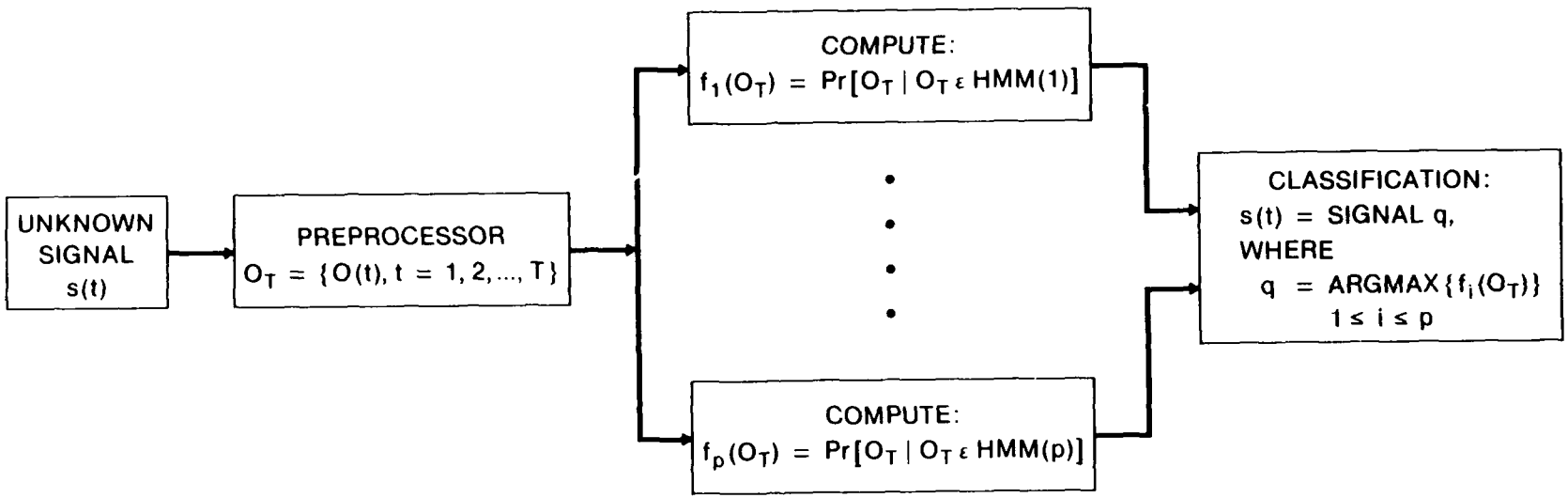


Figure 1. Classification of Unknown Signal s(t) as One of p Signals for Which Trained HMMs Are Available

really do represent the signals. Unfortunately, algorithms for calculating $F_{ij}(x)$ directly from the HMM parameters are not known. For later reference, note that $F_{ij}(x) \neq F_{ji}(x)$ in general.

The moments of $dF_{ij}(x)$ are defined by the Riemann-Stieltjes integral

$$M_{ij}(k,T) = \int_{-\infty}^{\infty} x^k \, dF_{ij}(x) , \qquad k = 0, 1, 2, \ldots . \tag{3}$$

If $F_{ij}(x)$ is differentiable with derivative $F'_{ij}(x)$, then the moments can be written equivalently as the Riemann integral

$$M_{ij}(k,T) = \int_{-\infty}^{\infty} x^k \, F'_{ij}(x) \, dx .$$

The moments depend on the length T of the observation sequence because $F_{ij}(x)$ depends on T, as seen from equation (2). They uniquely determine $dF_{ij}(x)$ when they are all finite and the characteristic function of $dF_{ij}(x)$ has a positive radius of convergence.[4] From equations (1) and (2), it is clear that $dF_{ij}(x) = 0$ for $x < 0$ and for $x > 1$. Thus, from equation (3),

$$M_{ij}(k,T) = \int_{0}^{1} x^k \, dF_{ij}(x) \leq 1 ,$$

so that all the moments are finite. The series

$$\Phi_{ij}(\omega) = \sum_{v=0}^{\infty} \frac{(i\omega)^v}{v!} M_{ij}(v,T)$$

for the characteristic function of $dF_{ij}(\omega)$ is absolutely convergent with an infinite radius of convergence because, for fixed $\omega_0 \neq 0$, each summand is bounded above in magnitude by $|\omega_0|^v/v!$ and thus the radius of

convergence must be at least as large as $\exp(|\omega_0|)$. Consequently, the moments of $dF_{ij}(x)$ uniquely determine $dF_{ij}(x)$ and, thereby, $F_{ij}(x)$.

This paper presents an algorithm for computing explicitly the moments of $dF_{ij}(x)$ up to any desired order directly from the given underlying parameters of the HMMs involved. The only essential assumption made is the usual one that the HMMs have a finite number of states. However, it is not required that all HMMs have the same number of states.

This paper also presents examples that compare the first two theoretical moments with simulation results. The examples are of independent interest because they exhibit important features of posterior likelihood classification based on ergodic and left-to-right HMMs that theoretical analysis alone would not show as easily or as quickly. These features are important because they indicate how the internal structure of HMMs impact the performance of the system depicted in figure 1.

II. THE MOMENT ALGORITHM

The reader is assumed to be familiar with such first principles of HMMs as given in reference 2. It is not, however, necessary to read this section to understand the examples provided in section III.

A. FINITE SYMBOL HMMs

Let HMM(υ) be a hidden Markov process with $n(\upsilon)$ states, $\upsilon = 1, \ldots$. Subscripted indices will always be written as functions of their subscripts (for instance, $n(\upsilon)$ is used instead of n_υ) to avoid the later use of subscripted subscripts. Let the state transition probability matrix of HMM(υ) be denoted as $A^\upsilon = [a^\upsilon_{i(\upsilon),j(\upsilon)}]$, for $i(\upsilon), j(\upsilon) = 1, \ldots, n(\upsilon)$. Let the initial state probability vector of HMM(υ) be denoted as $\pi^\upsilon = [\pi^\upsilon_{i(\upsilon)}]$, for $i(\upsilon) = 1, \ldots, n(\upsilon)$.

We first restrict attention to finite symbol HMMs; that is, we suppose that every observation sequence $O_T = \{O(t), t = 1, \ldots, T\}$ is such that

$$O(t) \in V = \{V_1, \ldots, V_m\} ,$$

where V is the set of all possible output symbols of the preprocessor. The true nature of the symbols in V is of no importance here. HMMs assume that O(t) is a random variable whose probability density function depends on the current state of the Markov process. Let the discrete probability density function for HMM(υ) when it is in state $i(\upsilon)$ be denoted as $B^\upsilon_{i(\upsilon)}$, for $i(\upsilon) = 1, \ldots, n(\upsilon)$. Thus, each $B^\upsilon_{i(\upsilon)}$ is a row vector of length m. Stacking these row vectors gives the $n(\upsilon)$-by-m symbol probability matrix

$$B^\upsilon = [b^\upsilon_{i(\upsilon),j(\upsilon)}] = \begin{bmatrix} B^\upsilon_1 \\ B^\upsilon_2 \\ \vdots \\ B^\upsilon_{n(\upsilon)} \end{bmatrix} .$$

Note that

$$b^{v}_{i(v)}(V_{j(v)}) = b^{v}_{i(v),j(v)} ,$$

where we define

$$b^{v}_{i(v)}(O(t)) = Pr[O(t) \mid HMM(v) \text{ and Markov state} = i(v)] .$$

The assumption that the training phase is completed means that the parameters $\lambda^{v} = (\pi^{v}, A^{v}, B^{v})$ are known.

For discrete symbol HMMs, the cumulative function $F_{ij}(x)$ is an increasing step function with a finite number of jump discontinuities. Let X_{ij} denote the set of all values of x for which $F_{ij}(x)$ is discontinuous. It follows that $dF_{ij}(x) = 0$ if x is not in X_{ij} and that, for x in X_{ij},

$$\begin{aligned} dF_{ij}(x) &= F_{ij}(x+) - F_{ij}(x-) \\ &= Pr[f_i(O_T) = x \mid O_T \in HMM(j)] . \end{aligned} \tag{4}$$

Substituting equation (4) into equation (3) gives

$$\begin{aligned} M_{ij}(k,T) &= \sum_{x \in X_{ij}} x^{k} \, Pr[f_i(O_T) = x \mid O_T \in HMM(j)] \\ &= \sum_{O_T} \{f_i(O_T)\}^{k} \, Pr[O_T \mid O_T \in HMM(j)] \\ &= \sum_{O_T} \{Pr[O_T \mid O_T \in HMM(i)]\}^{k} \, Pr[O_T \mid O_T \in HMM(j)] . \end{aligned} \tag{5}$$

It is clear from equation (5) that $M_{ij}(k,T) \neq M_{ji}(k,T)$ in general for $k > 1$. For $k = 1$, however, we have $M_{ij}(1,T) = M_{ji}(1,T)$ for all i, j, and T.

The expression in equation (5) is computable directly from the parameters of HMM(i) and HMM(j); however, such a calculation is not practical except for small T because the computational effort increases exponentially in T. To see this, note that the forward-backward algorithm[2] calculates $\Pr[O_T \mid O_T \in \text{HMM}(v)]$ using $n^2(v)\,T$ multiplications. Thus, each summand in equation (5) requires $k[n(i)\,n(j)]^2\,T^2$ multiplications. There are m^T different possible observation sequences $O_T = \{O(t),\ t = 1, \ldots, T\}$ because each O(t) can be any one of the m output symbols in V. Thus, direct calculation of equation (5) requires a total of $k[n(i)\,n(j)]^2\,T^2\,m^T$ multiplications.

We now derive a recursion for equation (5) that requires computational effort that grows only linearly with T. The recursion is derived for a more general expression that contains equation (5) as a special case. For k = 1, 2, ..., define

$$R(k,T) = \sum_{O_T} \prod_{v=1}^{k} \Pr[O_T \mid O_T \in \text{HMM}(v)] \ . \tag{6}$$

The application of equation (6) to compute any moment from equation (5) is straightforward; for example, R(k+1,T) equals $M_{21}(k,T)$ when HMM(2) = ... = HMM(k+1). Note that R(k,T) can be interpreted as a joint moment of HMMs.

The derivation of the recursion for R(k,T) proceeds as follows. The forward recursion portion of the forward-backward algorithm gives the expression

$$\Pr[O_T \mid O_T \in \text{HMM}(v)] = \sum_{j(v)=1}^{n(v)} \alpha_T^v(j(v)) \ , \tag{7}$$

where, for $2 \le t \le T$,

$$\alpha_t^v(j(v)) = \left[\sum_{i(v)=1}^{n(v)} \alpha_{t-1}^v(i(v))\ a_{i(v),j(v)}^v \right] b_{j(v)}^v(O(t)) \ , \tag{8}$$

and

$$\alpha_1^v(j(v)) = \pi_{j(v)}^v\ b_{j(v)}^v(O(1)) \ . \tag{9}$$

Substitute equation (7) into equation (6) to obtain

$$R(k,T) = \sum_{\substack{j(v)=1 \\ v=1,\ldots,k}}^{n(v)} \sum_{O_T} \prod_{v=1}^{k} \alpha_T^v(j(v))$$

$$= \sum_{\substack{j(v)=1 \\ v=1,\ldots,k}}^{n(v)} \mu_T(j(1), \ldots, j(k)) , \tag{10}$$

where we define for $t = 1, \ldots, T$

$$\mu_t(j(1), \ldots, j(k)) = \sum_{O_t} \prod_{v=1}^{k} \alpha_t^v(j(v)) . \tag{11}$$

One interpretation of μ_T is that it equals $R(k,T)$ given that $HMM(v)$ must end in state $j(v)$, $v=1, \ldots, k$. We seek a recursion for μ_T. First suppose that $2 \leq t \leq T$. Then, substituting equation (8) into equation (11) gives

$$\mu_t(j(1), \ldots, j(k)) = \sum_{O_t} \prod_{v=1}^{k} \left\{ \sum_{i(v)=1}^{n(v)} \alpha_{t-1}^v(i(v))\, a_{i(v),j(v)}^v\, b_{j(v)}^v(O(t)) \right\}$$

$$= \sum_{\substack{i(v)=1 \\ v=1,\ldots,k}}^{n(v)} \sum_{O_t} \left\{ \left[\prod_{v=1}^{k} \alpha_{t-1}^v(i(v)) \right] \left[\prod_{v=1}^{k} a_{i(v),j(v)}^v \right] \left[\prod_{v=1}^{k} b_{j(v)}^v(O(t)) \right] \right\}$$

$$= \sum_{\substack{i(v)=1 \\ v=1,\ldots,k}}^{n(v)} \left[\prod_{v=1}^{k} a_{i(v),j(v)}^v \right] \left\{ \sum_{O_t} \left[\prod_{v=1}^{k} \alpha_{t-1}^v(i(v)) \right] \left[\prod_{v=1}^{k} b_{j(v)}^v(O(t)) \right] \right\} .$$

Because $\alpha^{v}_{t-1}(i(v))$ does not depend on the last symbol $O(t)$ in the observation sequence $O_t = \{O(1), \ldots, O(t)\}$, we have

$$\mu_t(j(1), \ldots, j(k))$$

$$= \sum_{\substack{i(v)=1 \\ v=1,\ldots,k}}^{n(v)} \left[\prod_{v=1}^{k} a^{v}_{i(v),j(v)}\right] \left\{\sum_{O_{t-1}} \left[\prod_{v=1}^{k} \alpha^{v}_{t-1}(i(v))\right] \sum_{O(t)} \left[\prod_{v=1}^{k} b^{v}_{j(v)}(O(t))\right]\right\} .$$

Because the sum over $O(t)$ is independent of the observation sequence $O_{t-1} = \{O(1), \ldots, O(t-1)\}$, as well as the indices $i(v)$, and because of equation (11), we have

$$\mu_t(j(1), \ldots, j(k))$$

$$= \Gamma(j(1), \ldots, j(k)) \sum_{\substack{i(v)=1 \\ v=1,\ldots,k}}^{n(v)} \left[\prod_{v=1}^{k} a^{v}_{i(v),j(v)}\right] \mu_{t-1}(i(1), \ldots, i(k)) , \tag{12}$$

where

$$\Gamma(j(1), \ldots, j(k)) = \sum_{O(t)} \prod_{v=1}^{k} b^{v}_{j(v)}(O(t))$$

$$= \sum_{s=1}^{m} \prod_{v=1}^{k} b^{v}_{j(v)}(V_s) . \tag{13}$$

Equation (12) is the desired recursion for $2 \leq t \leq T$. For $t = 1$, substituting equation (9) into equation (11) gives

$$\mu_1(j(1), \ldots, j(k)) = \sum_{O(1)} \prod_{v=1}^{k} \alpha^{v}_{1}(j(v))$$

$$= \left(\prod_{v=1}^{k} \pi_{j(v)}^{v} \right) \sum_{O(1)} \prod_{v=1}^{k} b_{j(v)}^{v}(O(1))$$

$$= \Gamma(j(1), \ldots, j(k)) \prod_{v=1}^{k} \pi_{j(v)}^{v} \cdot \qquad (14)$$

Let $k = 1$. From the definition, it is clear that $R(1,T) = 1$ for all T, regardless of HMM(1), because the sum in equation (6) is over all O_T. To independently check the recursion (12)-(13), note that, from equation (13),

$$\Gamma(j(1)) = \sum_{s=1}^{m} b_{j(1)}^{1}(V_s) = 1 , \qquad 1 \leq t \leq T .$$

From equation (14), we have

$$\mu_1(j(1)) = \pi_{j(1)}^{1} .$$

Hence, from equation (10), we obtain

$$R(1,1) = \sum_{j(1)=1}^{n(1)} \pi_{j(1)}^{1} = 1 .$$

The recursion is verified for $T = 1$. For $T = 2$, from equation (12), we have

$$\mu_2(j(1)) = \sum_{i(1)=1}^{n(1)} \mu_1(i(1))\, a_{i(1),j(1)}^{1}$$

$$= \sum_{i(1)=1}^{n(1)} \pi_{i(1)}^{1}\, a_{i(1),j(1)}^{1}$$

so that, from equation (10),

$$R(1,2) = \sum_{j(1)=1}^{n(1)} \left\{ \sum_{i(1)=1}^{n(1)} \pi^1_{i(1)} \, a^1_{i(1),j(1)} \right\}$$

$$= \sum_{i(1)=1}^{n(1)} \left\{ \pi^1_{i(1)} \sum_{j(1)=1}^{n(1)} a^1_{i(1),j(1)} \right\}$$

$$= 1 \ ,$$

and the recursion is verified for T = 2.

The first nontrivial special case is k = 2. In this case, R(2,T) is identically the first moment $M_{12}(1,T)$. From equation (12), we have for $2 \leq t \leq T$

$$\mu_t(j(1),j(2)) = \Gamma(j(1),j(2)) \sum_{i(1)=1}^{n(1)} \sum_{i(2)=1}^{n(2)} \mu_{t-1}(i(1),i(2)) \, a^1_{i(1),j(1)} \, a^2_{i(2),j(2)} \ ,$$

and, from equation (14),

$$\mu_1(j(1),j(2)) = \Gamma(j(1),j(2)) \, \pi^1_{j(1)} \, \pi^2_{j(2)}$$

where, from equation (13),

$$\Gamma(j(1),j(2)) = \sum_{s=1}^{m} b^1_{j(1)}(V_s) \, b^2_{j(2)}(V_s) \ .$$

From equation (10), then, we have

$$R(2,T) = \sum_{j(1)=1}^{n(1)} \sum_{j(2)=1}^{n(2)} \mu_T(j(1),j(2)) \ .$$

Computation of $R(2,T) = M_{12}(1,T)$ is therefore not excessively laborious.

The evaluation of $R(k,T)$ using the recursion (12) is properly broken into two parts. The first is the precalculation of $\Gamma(j(1), \ldots, j(k))$ for every possible value of the indices $j(\upsilon)$. This requires $(k-1)m\ N^k$ multiplications and N^k storage locations, where

$$N = \left[\sum_{\upsilon=1}^{k} n(\upsilon)\right]^{1/k} \tag{15}$$

is the geometric mean of the number of different states in the various HMMs and is not necessarily an integer. If $N = 8$ and if there are $m = 16$ different observation symbols, then computing and storing Γ for $k = 6$ requires 262144 storage locations and 2.1×10^7 multiplications. Storage is clearly more crucial an issue than multiplications.

It is possible in some cases to utilize the underlying symmetries of Γ to reduce both storage and computational effort. For example, if $HMM(2) = \ldots = HMM(k+1)$, then

$$\Gamma(j(1),j(2), \ldots, j(k+1)) = \Gamma(j(1), \sigma(j(2)), \ldots, \sigma(j(k+1))) \tag{16}$$

for every permutation σ of the k integers $j(2), \ldots, j(k+1)$. The proof of equation (16) follows easily from equation (13) because multiplication is commutative. Thus one only need consider indices that satisfy

$$1 \le j(1) \le n(1) \text{ and } 1 \le j(2) \le j(3) \le \ldots \le j(k+1) \le n(2) \ .$$

The number of ordered index sets $\{j(\upsilon)\}$ is equal to the number of combinations of $n(2)$ letters taken k at a time, when each letter may be repeated any number of times up to k. Storage is therefore proportional to

$$N_{k+1} = \left(\frac{n(2)(n(2)+1) \ldots (n(2)+k-1)}{k!}\right) n(1) \ ,$$

which is significantly smaller than the $[n(2)]^k\, n(1)$ storage that would otherwise be necessary. The total multiplication count is also reduced proportionately.

Once Γ has been computed and stored for a given value of k, the recursion (12) can be computed for any length T of the observation sequence. For each of the N^k sets of indices $\{j(\upsilon)\}$ in equation (12), the sum over all N^k indices $\{i(\upsilon)\}$ must be undertaken. This sum appears to require $k\, N^k$ multiplications; however, by using the nested form,

$$\sum_{i(1)=1}^{n(1)} a^1_{i(1),j(1)} \left[\sum_{i(2)=1}^{n(2)} \cdots \left[\sum_{i(k)=1}^{n(k)} a^k_{i(k),j(k)}\, \mu_{t-1}(i(1), \ldots, i(k)) \right] \cdots \right],$$

it is possible to use approximately

$$N^k + N^{k-1} + \ldots + N^2 + N = \left(\frac{N}{N-1}\right)(N^k - 1)$$

instead. If lower order terms are neglected, computing one iteration of equation (12) requires about N^{2k} multiplications. For an observation sequence of length T, computing μ_T requires on the order of $N^{2k}T$ multiplications. If $N = 8$ and $T = 32$, then 2.2×10^{12} multiplications are required for $k = 6$. Assuming a multiplication takes one microsecond, the calculation requires 611 hours and is clearly impractical.

Significant reduction in computational effort is possible in some cases by utilizing the underlying symmetries in μ_t. For example, if HMM(2) = ... = HMM(k+1), then

$$\mu_t(j(1), j(2), \ldots, j(k+1)) = \mu_t(j(1), \sigma(j(2)), \ldots, \sigma(j(k+1))) \qquad (17)$$

for every permutation σ of the k integers j(2), ..., j(k+1). The proof of equation (17) follows easily by induction from equation (12) because multiplication is commutative and because Γ satisfies the same symmetry property (16) in this case. Thus, the recursion (12) need be computed for

only N_{k+1} sets of indices. The total multiplication count is reduced to $4N_{k+1}^2 T$, which is significantly smaller than the $N^{2k}T$ multiplications that would otherwise be needed. For the above example requiring 611 hours, if $N = n(1) = n(2) = 8$ and if the symmetry (17) is utilized, the calculation would be reduced to roughly a 96-minute calculation. Utilizing symmetry is clearly significant in that it can turn an impractical long calculation into a feasible shorter one.

Underflow is potentially a problem when the recursion (12) is computed. It can be easily overcome in exactly the same manner as pointed out in reference 2 for preventing numerical underflow during the calculation of the forward-backward algorithm. Specifically, let μ_t be computed according to equation (12) and then multiplied by a scale factor c_t defined by

$$c_t = \left[\sum_{\substack{j(v)=1 \\ v=1,\ldots,k}}^{n(v)} \mu_t(j(1), \ldots, j(k)) \right]^{-1} . \tag{18}$$

Then use the scaled values of μ_t in the recursion (12) to compute μ_{t+1}, which is in turn scaled as shown in equation (18). If we continue in this fashion and recall the expression in equation (10), it follows that

$$R(k,T) = \left(\prod_{t=1}^{T} c_t \right)^{-1} . \tag{19}$$

Because the product cannot be evaluated without underflow, we compute instead

$$\log R(k,T) = -\sum_{t=1}^{T} \log c_t . \tag{20}$$

Any convenient scale factor can be used instead of equation (18). A potentially useful one might be to take $\bar{c}_t = N^k$. Using $\bar{c}_t$ would eliminate the effort of computing the sum in equation (18) before scaling.

B. CONTINUOUS SYMBOL HMMS

The objective of this section is to show that the moment algorithm for discrete symbol HMMs can be carried over essentially unchanged to continuous symbol HMMs. In fact, it holds also for continuous vector symbol HMMs; however, only the continuous symbol HMMs are treated here for simplicity.

Throughout this section, it is assumed that each output symbol $O(t)$ is a real random variable defined on some underlying event space, V. The probability density function of $O(t)$ is uniquely defined for each state $i(\upsilon) = 1, \ldots, n(\upsilon)$ of each $HMM(\upsilon)$, $\upsilon = 1, 2, \ldots$, and is denoted as $b^{\upsilon}_{i(\upsilon)}(x)$. Thus, for real numbers α and β with $\alpha < \beta$, we have

$$\int_{\alpha}^{\beta} b^{\upsilon}_{i(\upsilon)}(x)\, dx = \Pr[\alpha \leq O(t) \leq \beta \mid \text{HMM}(\upsilon) \text{ and Markov state} = i(\upsilon)] \,. \quad (21)$$

An *observation sequence* $O_T = \{x_t, t = 1, 2, \ldots, T\}$ is a sequence of real numbers x_t, with x_t being a realization of the random variable $O(t)$. The posterior likelihood function $f_{\upsilon}(O_T)$ is now a probability density function for continuous symbol HMMs, as opposed to a simple probability (see equation (1)) for discrete symbol HMMs. Thus, for real vectors $\vec{\alpha}$ and $\vec{\beta}$ with $\vec{\alpha} < \vec{\beta}$, we have

$$\int_{\vec{\alpha}}^{\vec{\beta}} f_{\upsilon}(O_T)\, dO_T = \Pr[\vec{\alpha} \leq O_T \leq \vec{\beta} \mid O_T \in \text{HMM}(\upsilon)] \,, \quad (22)$$

where $O_T \in \text{HMM}(\upsilon)$ denotes the hypothesis that O_T is a realization of $\text{HMM}(\upsilon)$ and $dO_T = dx_1 \ldots dx_T$.

The conditional cumulative distribution functions $F_{ij}(x)$ are defined by equation (2), just as for discrete symbol HMMs. From equation (3), we have the moments

$$M_{ij}(k,T) = \int_{-\infty}^{\infty} x^k \, dF_{ij}(x)$$

$$= \int_{-\infty}^{\infty} x^k \, \Pr[x \le f_i(O_T) \le x + dx \mid O_T \in \text{HMM}(j)] \, dx$$

$$= \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{\text{T-fold}} \{f_i(O_T)\}^k \, f_j(O_T) \, dO_T \, , \tag{23}$$

which is the continuous analog of equation (5). It is clear from equation (23) that $M_{ij}(k,T) = M_{ji}(k,T)$ in general only for the special case $k = 1$. The analog of equation (6) for continuous HMMs is

$$R(k,T) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{\text{T-fold}} \prod_{v=1}^{k} f_v(O_T) \, dO_T \, . \tag{24}$$

The forward-backward algorithm for computing the posterior likelihood function for continuous HMMs is modified[5] as follows:

$$f_v(O_T) = \sum_{j(v)=1}^{n(v)} \alpha_T^v(j(v)) \, , \tag{25}$$

where $\alpha_T^v(j(v))$ is computed exactly as given by the recursions (8) and (9), with the only difference being that $b_{j(v)}^v(O(t))$ in equation (8) is now interpreted as the probability density function implicit in equation (21). Consequently, equation (10) still holds exactly if we define

$$\mu_t(j(1), \ldots, j(k)) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{t\text{-fold}} \prod_{v=1}^{k} \alpha_t^v(j(v))\, dO_t \tag{26}$$

as the analog of equation (11). Proceeding as before with t-fold integrals replacing t-fold summations gives exactly the recursion (12), but with the one-dimensional integral

$$\Gamma(j(1), \ldots, j(k)) = \int_{-\infty}^{\infty} \prod_{v=1}^{k} b_{j(v)}^v(x)\, dx \tag{27}$$

in place of equation (13).

The remarks in the preceding section concerning storage, multiplication counts, and symmetry properties all apply for continuous symbol HMMs. The primary difference is that equation (27) requires an integral evaluation instead of a finite sum as in equation (13). This evaluation increases the initial computational overhead, but once equation (27) is computed, the algorithm (12) proceeds exactly as before.

III. COMPARISON OF THEORETICAL MOMENTS WITH SIMULATION

Ergodic Markov chains are those for which it is possible to transition from every state to every other state, although not necessarily in one step. Left-to-right Markov chains are those for which transitions to lower numbered states are not allowed, that is, have probability zero. These two types of chains are sufficiently different that they are considered separately in the examples.

One interpretation is that ergodic HMMs are models of quasi-stationary signals, while left-to-right HMMs are models of transient signals that ultimately become stationary (because the highest numbered state is not exited once it is entered). One might therefore expect these two types of HMMs to impact the performance of the maximum likelihood classifier depicted in figure 1 in different ways. The three examples given in this section support this expectation.

Using the above interpretation, the examples may be described as follows. The first example shows that maximum likelihood classification based on HMMs can reliably distinguish between sufficiently long quasi-stationary signals with a reasonable amount of computational effort. The second example shows that short quasi-stationary and transient signals look significantly different to the HMM transient recognizer, but <u>not</u> to the HMM recognizer based on the quasi-stationary signal. The third example shows that noisy observations of transient signals adversely affect classification performance by making the transient signal appear to have a stationary component, which is then misclassified by the HMM transient recognizer.

A. TWO ERGODIC HMMS

HMM(1) and HMM(2) are five-state, eight-symbol ergodic models whose parameters are given (rounded to three significant decimals) in tables 1 and 2, respectively. HMM(1) clearly generates observation sequences of uniformly distributed symbols. HMM(2) is more complex in structure, but every symbol can be generated in every state. The fundamental question of interest here is the following. How long must an observation sequence be to

guarantee that maximum posterior probability classification (described in the Introduction) will be 98 percent reliable? We will give what may best be described as a semiempirical answer to this question.

Because of the nature of HMM(1), it is easy to see that

$$f_1(O_T) = \Pr[O_T \mid O_T \in \text{HMM}(1)] = 8^{-T} .$$

In other words, the posterior likelihood function based on HMM(1) is constant because all observation sequences are equally likely if $O_T \in$ HMM(1). In particular, $f_1(O_T)$ cannot distinguish $O_T \in$ HMM(1) from $O_T \in$ HMM(2) and thus is useless for classification.

The posterior likelihood function based on HMM(2), instead of HMM(1), is useful for classification. Ten-thousand observation sequences O_T of each HMM were generated, and the posterior probability $f_2(O_T)$ was computed

Table 1. Parameters of HMM(1)

NUMBER OF MARKOV STATES = 5

NUMBER OF SYMBOLS PER STATE = 8

INITIAL STATE PROBABILITY VECTOR:

2.00E-01	2.00E-01	2.00E-01	2.00E-01	2.00E-01

TRANSITION PROBABILITY MATRIX:

2.00E-01	2.00E-01	2.00E-01	2.00E-01	2.00E-01
2.00E-01	2.00E-01	2.00E-01	2.00E-01	2.00E-01
2.00E-01	2.00E-01	2.00E-01	2.00E-01	2.00E-01
2.00E-01	2.00E-01	2.00E-01	2.00E-01	2.00E-01
2.00E-01	2.00E-01	2.00E-01	2.00E-01	2.00E-01

SYMBOL PROBABILITY MATRIX (TRANSPOSED):

1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01
1.25E-01	1.25E-01	1.25E-01	1.25E-01	1.25E-01

Table 2. Parameters of HMM(2), Rounded to Three Significant Digits

NUMBER OF MARKOV STATES = 5

NUMBER OF SYMBOLS PER STATE = 8

INITIAL STATE PROBABILITY VECTOR:

1.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00

TRANSITION PROBABILITY MATRIX:

1.40E-01	2.35E-01	3.08E-01	1.24E-01	1.94E-01
1.40E-01	1.14E-01	2.99E-01	2.13E-01	2.34E-01
4.37E-02	3.20E-01	1.72E-01	1.27E-01	3.38E-01
9.73E-02	4.97E-01	1.53E-02	1.15E-01	2.75E-01
2.36E-01	2.49E-02	4.27E-01	2.82E-01	2.98E-02

SYMBOL PROBABILITY MATRIX (TRANSPOSED):

1.81E-01	1.22E-01	7.89E-03	1.48E-01	7.04E-02
1.39E-01	8.28E-02	3.23E-02	9.13E-02	1.33E-01
2.67E-02	1.60E-01	5.87E-02	1.08E-01	2.34E-01
1.79E-01	1.66E-01	2.18E-01	1.30E-01	5.97E-02
1.56E-01	1.58E-01	2.15E-01	2.09E-01	2.35E-01
1.19E-01	5.75E-02	1.11E-01	1.02E-01	1.03E-01
1.76E-01	1.32E-01	2.40E-01	6.61E-02	1.76E-02
2.37E-02	1.22E-01	1.17E-01	1.46E-01	1.47E-01

using the forward-backward algorithm. Figure 2 shows a histogram of the natural logarithm of $dF_{22}(x)$ for $T = 25$. The observation sequences are thus matched to the posterior likelihood function. Figure 3 shows a histogram of $\log dF_{21}(x)$ for $T = 25$. In figure 3, then, O_1 is mismatched to the likelihood function. As is clear from figures 2 and 3, the difference between the mean values of the log likelihood functions is about 1.4 standard deviations. Thus, the potential exists for using $\log dF_{22}(x)$ to classify observation sequences; however, $T = 25$ is not long enough to classify with high reliability.

A useful observation drawn from figures 2 and 3 is that the probability density function of $\log dF_{2j}(x)$ is nicely approximated by the normal distribution. Let μ_{ij} and σ_{ij} denote the mean and standard deviation of $\log dF_{ij}(x)$. Then, if $dF_{ij}(x)$ is log-normal, it is easy to show that μ_{ij} and σ_{ij} are related to the moments $M_{ij}(k,T)$ by the formulas

$$\mu_{ij} = 2 \log M_{ij}(1,T) - (1/2) \log M_{ij}(2,T) \tag{28}$$

$$\sigma_{ij}^2 = \log M_{ij}(2,T) - 2 \log M_{ij}(1,T) \; . \tag{29}$$

It is stressed that equations (28) and (29) hold exactly if and only if $dF_{ij}(x)$ is truly log-normal. For finite symbol HMMs, $dF_{ij}(x)$ is necessarily discrete, so that both equations (28) and (29) must be viewed as approximations. Sufficient conditions under which it may be proved that $dF_{ij}(x)$ is, in some sense, approximately log-normal are unknown.

Table 3 gives a comparison between the mean and standard deviations of $\log dF_{ij}(x)$ estimated from 10000 observation sequences O_T and those calculated from equations (28) and (29). This table shows good agreement between the approximations of equations (28) and (29) and the sample means and variances. It also establishes that observation sequences of length $T \approx 400$ are long enough to distinguish between $O_T \in$ HMM(1) and $O_T \in$ HMM(2) with high reliability. That is, the difference between the mean value of $\log dF_{21}(x)$ and the mean value of $\log dF_{22}(x)$ is about 5.2 standard deviations. Assuming $\log dF_{21}(x)$ and $\log dF_{22}(x)$ are normally distributed, as they appear to be, then the reliability of the maximum posterior classifier is about 98 percent.

Computing the posterior likelihood function $f_2(O_T)$ for $T = 400$ requires $n^2T = 10000$ multiplications. This is about the same level of effort as computing one 512-point FFT, which requires $(2)(512)(\log_2 512) = 9216$ multiplications. In other words, the computational requirements of $f_2(O_{400})$ are small enough for practical application. Furthermore, the forward-backward algorithm for computing $f_2(O_T)$ is mathematically equivalent to a nested sequence of matrix-vector multiplications. Consequently, it is possible to reduce total computation time by the design of a "black box" to exploit this special structure in hardware.

B. MIXED ERGODIC AND LEFT-TO-RIGHT HMMS

HMM(3) is a five-state, eight-symbol left-to-right model whose parameters are given in table 4. It has a structure that might conceivably

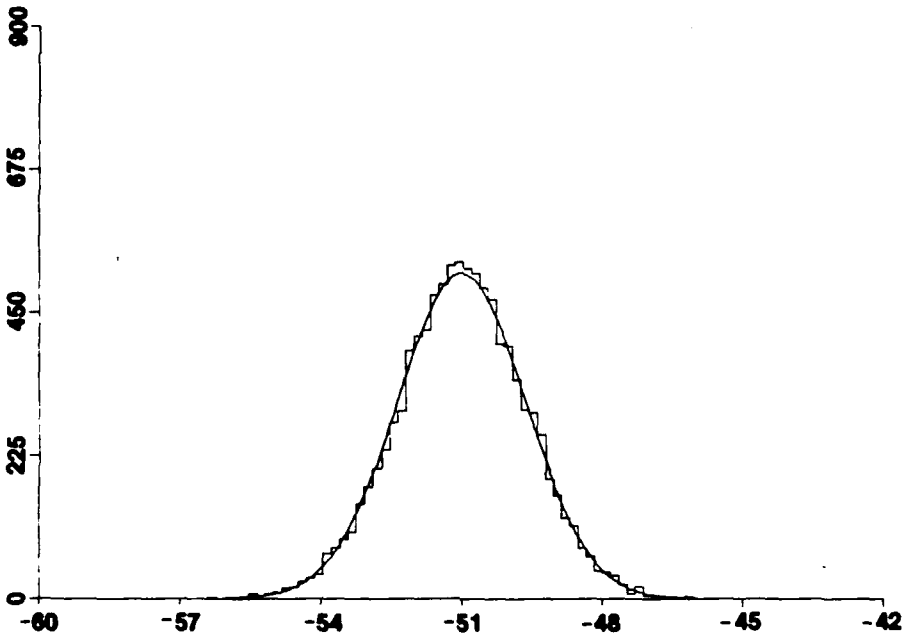


Figure 2. Histogram of 10000 Values of log $dF_{22}(x)$ for T = 25.
(The Normal Curve Has the Sample Mean and Variance Given in Table 3.)

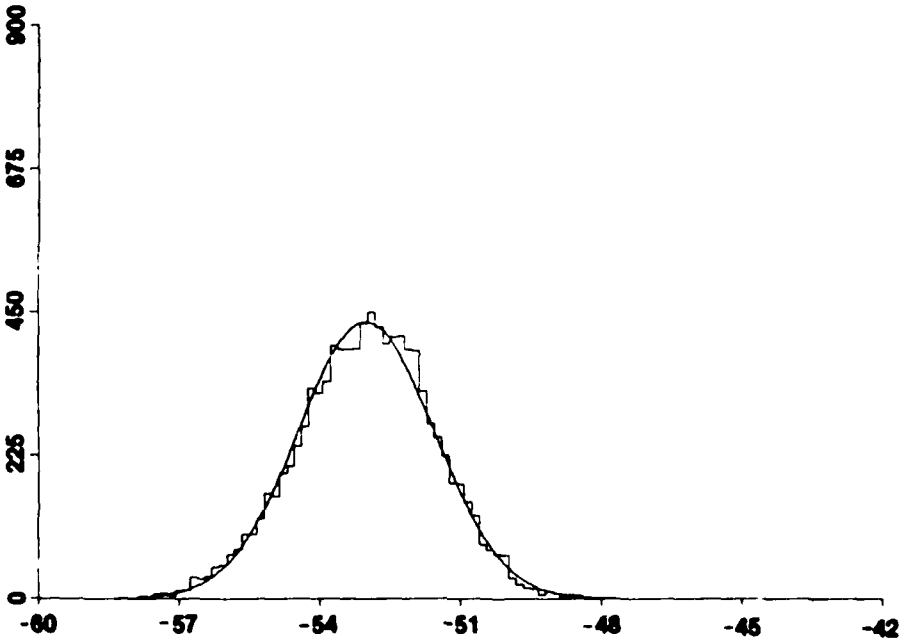


Figure 3. Histogram of 10000 Values of log $dF_{21}(x)$ for T = 25.
(The Normal Curve Has the Sample Mean and Variance Given in Table 3.)

Table 3. Comparison of Two Estimates for the Mean and Standard Deviation of log $dF_{2j}(x)$ for j = 1, 2

	T	Mean Value		Standard Deviation	
		Sample	Eq. 28	Sample	Eq. 29
j = 1	5	-10.8	-10.6	0.95	0.71
	10	-21.3	-21.2	1.11	0.92
	15	-31.9	-31.8	1.24	1.10
	20	-42.4	-42.4	1.35	1.25
	25	-53.0	-52.9	1.47	1.38
	50	-105.8	-105.8	1.93	1.91
	100	-211.4	-211.5	2.62	2.67
	200	-422.6	-423.0	3.60	3.76
	400	-845.0	-845.8	5.09	5.30
j = 2	5	-10.1	-10.1	0.69	0.59
	10	-20.3	-20.3	0.90	0.84
	15	-30.6	-30.5	1.08	1.03
	20	-40.8	-40.8	1.23	1.20
	25	-51.0	-51.0	1.37	1.34
	50	-102.1	-102.1	1.92	1.90
	100	-204.4	-204.4	2.66	2.69
	200	-408.9	-409.0	3.77	3.80
	400	-818.0	-818.1	5.33	5.37

arise in the SIIWR problem. Note that HMM(3) never leaves the fifth state once it is entered. Consequently, all sufficiently long observation sequences ultimately contain only the three symbols V_6, V_7, and V_8. Note also that the symbol V_8 occurs if and only if the fifth state has been entered. It follows that an observation sequence O_T containing the symbol V_8 and subsequently containing any of the five symbols V_1, V_2, V_3, V_4, or V_5 must have posterior likelihood zero; i.e., $f_3(O_T) = 0$. Other forbidden symbol sequences may also be noticed. It will be seen that these facts make $f_3(O_T)$ a powerful discriminator against ergodic observation sequences. To summarize briefly, this example will show that short observation sequences of quasi-stationary and transient HMMs look very different to the transient HMM recognizer. On the other hand, all observation sequences look somewhat alike to ergodic HMM recognizers.

When HMM(3) enters its fifth state, it becomes stationary and,

Table 4. Parameters of HMM(3)

NUMBER OF MARKOV STATES = 5

NUMBER OF SYMBOLS PER STATE = 8

INITIAL STATE PROBABILITY VECTOR:

1.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00

TRANSITION PROBABILITY MATRIX:

6.00E-01	4.00E-01	0.00E+00	0.00E+00	0.00E+00
0.00E+00	7.00E-01	2.00E-01	1.00E-01	0.00E+00
0.00E+00	0.00E+00	6.00E-01	4.00E-01	0.00E+00
0.00E+00	0.00E+00	0.00E+00	7.00E-01	3.00E-01
0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E+00

SYMBOL PROBABILITY MATRIX (TRANSPOSED):

9.00E-01	1.00E-01	0.00E+00	0.00E+00	0.00E+00
1.00E-01	6.00E-01	0.00E+00	0.00E+00	0.00E+00
0.00E+00	2.00E-01	3.00E-01	0.00E+00	0.00E+00
0.00E+00	1.00E-01	6.00E-01	1.00E-01	0.00E+00
0.00E+00	0.00E+00	1.00E-01	2.00E-01	0.00E+00
0.00E+00	0.00E+00	0.00E+00	4.00E-01	1.00E-01
0.00E+00	0.00E+00	0.00E+00	3.00E-01	6.00E-01
0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.00E-01

consequently, significantly less interesting. Insight into the length of the transient portion of HMM(3) observation sequences is gained by estimating the first passage time of HMM(3) into its fifth state, that is, the number of transitions in the Markov process before its fifth state is entered. *The mean and variance of first passage times may be computed explicitly;*[6] however, simulation was used here instead. In 10000 observation sequences generated for HMM(3), it was found that the mean and standard deviation of the first passage time was 10.9 and 4.8, respectively. The least first passage time was 3 transitions, and the largest first passage time was 43 transitions. Thus, observation sequences almost certainly become stationary for $t \geq 50$.

Figure 4 and table 5 clearly show that $dF_{33}(x)$ is a "well-behaved" distribution even though HMM(3) is not ergodic. However, $dF_{33}(x)$ is not as closely approximated by a log-normal distribution as are $dF_{21}(x)$ and $dF_{22}(x)$, as evidenced by the discrepancy in table 5 between the sample

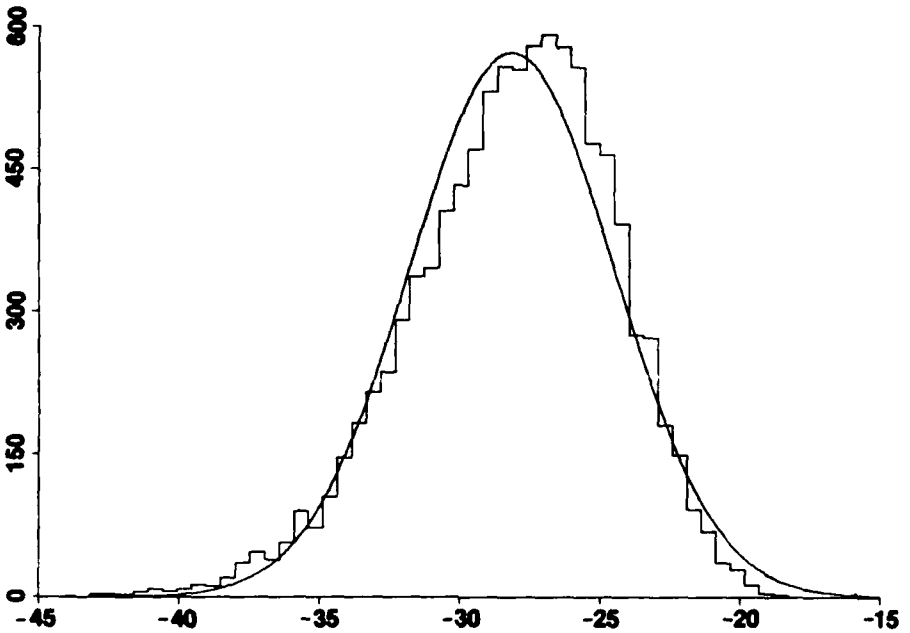


Figure 4. Histogram of 10000 Values of log $dF_{33}(x)$ for T = 25.
(The Normal Curve Has the Sample Mean and Variance Given in Table 5.)

statistics and the statistics that would hold if $dF_{33}(x)$ were truly log-normal.

Ten-thousand observation sequences of HMM(1) and HMM(2) were generated and the posterior likelihood $f_3(O_T)$ was computed using the forward-backward algorithm. The observation sequences are thus mismatched to the posterior likelihood function. Table 6 gives the number of sequences for which $f_3(O_T) = 0$. Better than 99-percent rejection of the simulated ergodic HMM observations was attained when T = 10, that is, when the observation sequences were about as long as the mean first passage time of HMM(3) into state 5. Total rejection of the 10000 ergodic observations occurred for T = 20.

The ability of $f_3(O_T)$ to reject observations of $O_T \in$ HMM(2) is much more impressive than the $f_2(O_T)$ rejection of $O_T \in$ HMM(3). The lack of symmetry $F_{ij}(x) \neq F_{ji}(x)$ is striking in this instance. Table 7 gives two estimates of the mean and standard deviation of log $dF_{23}(x)$, and figure 5 is a histogram of the case T = 25. The mean values of the 10000 samples and those predicted by equation (28) agree very well; however,

Table 5. Comparison of Two Estimates for the Mean and Standard Deviation of log $dF_{33}(x)$

T	Mean Value		Standard Deviation	
	Sample	Eq. 28	Sample	Eq. 29
5	- 5.6	- 4.9	1.92	1.13
10	-12.8	-11.8	2.30	1.91
15	-18.6	-18.2	2.72	2.33
20	-23.5	-22.6	3.22	2.29
25	-28.1	-26.3	3.61	2.15
50	-50.5	-47.2	4.59	2.75

Table 6. Number of $O_T \in HMM(i)$ for Which $f_3(O_T) = 0$, $i = 1, 2$

T	HMM(1)	HMM(2)
5	9389	9172
10	9937	9918
15	9997	9988
20	10000	10000

Table 7. Comparison of Two Estimates for the Mean and Standard Deviation of log $dF_{23}(x)$

T	Mean Value		Standard Deviation	
	Sample	Eq. 28	Sample	Eq. 29
5	- 10.8	-10.8	0.51	0.60
10	- 21.4	-21.4	0.91	0.89
15	- 32.0	-32.0	1.02	1.04
20	- 42.6	-42.7	1.04	1.15
25	- 53.2	-53.5	1.05	1.12
50	-106.4	-106.8	1.11	1.43

$dF_{23}(x)$ is not as well approximated by a log-normal as $dF_{22}(x)$ and $dF_{11}(x)$, as seen from the discrepancy in the sample versus the predicted standard deviations. In any event, it is clear by comparing table 7 with

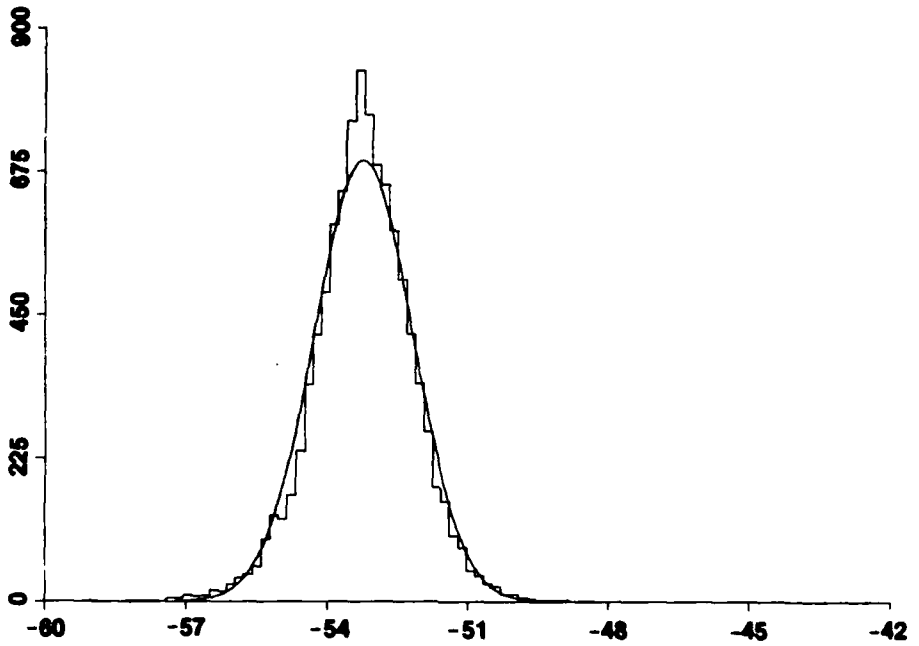


Figure 5. Histogram of 10000 Values of log $dF_{23}(x)$ for T = 25. (The Normal Curve Has the Sample Mean and Variance Given in Table 7.)

the lower half of table 3 that $f_2(O_T)$ cannot reliably distinguish $O_T \in$ HMM(3) from $O_T \in$ HMM(2) when T = 50. However, since the first passage time of HMM(3) is almost certainly less than T = 50, increasing the observation sequence length to improve reliability is not appropriate if the underlying intent is the classification of the transient portion of HMM(3).

C. LEFT-TO-RIGHT HMM WITH NOISE

In this example, the effect of noise on the maximum posterior likelihood classifier is assessed for the left-to-right model HMM(3). The right way to study noise in finite symbol HMMs is to add the noise to the original time series s(t) and then analyze the particular preprocessor under consideration to determine the noisy symbol sequence. However, no particular preprocessor is proposed here, and so we resort to modeling noise in much the same way that Shannon modeled noisy discrete memoryless channels.[7] This approach can give an indication of the successful classification rate as a function of the probable number of incorrect symbols in an observation sequence, but it cannot provide an assessment of the effect of signal-to-noise ratio on

classification because such an assessment requires knowledge of the preprocessor.

Denote by h_{kj} the probability that the observation symbol V_k is altered to symbol V_j by the noise mechanism and define the m-by-m noise probability matrix $H = [h_{kj}]$. It is assumed that H is independent of the state of the Markov chain and of time t. Consequently, the output of a given HMM corrupted by noise is equivalent to another HMM that is noiseless. If $\lambda = (\pi, A, B)$ are the parameters of a given HMM with noise matrix H, the parameters of the equivalent noiseless HMM are $\tilde{\lambda} = (\pi, A, BH)$. The proof is straightforward: the product $b_{ik} h_{kj}$ is the probability that the state of the Markov chain is i and that symbol j is produced, given that symbol k was the output of the given HMM. The sum over k of $b_{ik} h_{kj}$ gives the component $\tilde{b}_{ij}$ of the equivalent noiseless HMM symbol probability matrix $\tilde{B}$. Clearly, $\tilde{b}_{ij}$ equals the (i,j) component of the product BH, so that $\tilde{B} = BH$.

The noise probability matrix H must be row stochastic; that is, every row sum must equal one. The HMM-generated symbol V_k is altered by noise to one of the available symbols, so that row k must sum to one.

Because H has row sums equal to one, the matrix $\tilde{B}$ is a valid symbol probability matrix for the equivalent noiseless HMM; that is, each row of $\tilde{B} = BH$ sums to one. We have

$$\sum_{j=1}^{m} \tilde{b}_{ij} = \sum_{j=1}^{m} \sum_{k=1}^{m} b_{ik} h_{kj}$$

$$= \sum_{k=1}^{m} b_{ik} \sum_{j=1}^{m} h_{kj}$$

$$= \sum_{k=1}^{m} b_{ik}$$

$$= 1 \quad .$$

The worst case noise probability matrix, denoted H^0, has the constant entry $h^0_{ij} = 1/m$ for all i and j. In this case,

$$\tilde{b}_{ij} = \sum_{k=1}^{m} b_{ik}\, h_{kj} = \frac{1}{m} \sum_{k=1}^{m} b_{ik} = \frac{1}{m} .$$

Consequently, HMMs with noise probability matrix H^0 are indistinguishable. In fact, H^0 makes all HMMs statistically equivalent to the ergodic HMM(1) given in table 1.

Let $Pr[V_i]$ be the relative frequency of occurrence of the symbol V_i in observation sequences of length T before the addition of noise. Thus, we have $\Sigma\ Pr[V_i] = 1$. After alteration by noise, the probability of correct occurrences of V_i in O_T is then $Pr[V_i]\ h_{ii}$. The probability that the symbol $O(t) \in O_T$ is correct is

$$D_T = \sum_{i=1}^{m} Pr[V_i]\ h_{ii} \tag{30}$$

and the probability that O(t) is incorrect is

$$E_T = 1 - D_T . \tag{31}$$

For the examples here, given a specific value of E_T, we choose the simple noise probability matrix H defined by

$$\begin{aligned} h_{ii} &= 1 - E_T , \qquad \text{all } i , \\ h_{ij} &= \frac{E_T}{m - 1} , \qquad \text{all } i \neq j . \end{aligned} \tag{32}$$

For this choice of H, D_T is independent of the actual values of $Pr[V_i]$, as is clear from equation (30) and the fact that $\Sigma\ Pr[V_i] = 1$.

Noise tends to make observations of all HMMs look like observations of HMM(1), and ergodic observation sequences tend to have forbidden symbol sequences for the left-to-right HMM(3). The first natural issue is therefore to determine how many forbidden symbol sequences occur as a function of the incorrect symbol probability E_T. Table 8 gives the results for various values of T and E_T, based on simulations of 10000 observation sequences. It shows that forbidden symbol sequences are less likely for small T than for large T. This table also shows that noisy observations of HMM(3) do not have as high a proportion of forbidden symbol sequences as observations of HMM(1) and HMM(2), even for E_T = 10 percent, as can be seen by comparing tables 6 and 8. One may conclude from table 8 that E_T must be small and T must be short to minimize misclassification due to forbidden symbol sequences. For instance, if T = 25 and E_T = 0.001, the misclassification rate is apparently at least 1.21 percent. Shorter T, however, causes smaller shifts in the statistics in the likelihood function and thus increases the misclassification rate. Consequently, a tradeoff exists between short T and long T.

The total misclassification rate can be expressed as the sum of the misclassification rate due to forbidden symbol sequences and the misclassification rate due to noise-induced shift in the statistics of the nonzero values of the posterior likelihood function. We examine the total misclassification rate for HMM(4), which is defined to be the HMM equivalent

Table 8. Number of $O_T \in$ HMM(3) + Noise for Which $f_3(O_T) = 0$ at Various Values of E_T

T	E_T = 0.1	0.01	0.001	0.0001
5	2194	236	23	1
10	3906	443	37	1
15	5305	651	64	11
20	6625	986	103	13
25	7643	1303	121	11
50	9643	2684	345	34

to HMM(3) with the noise matrix H given by equation (32) with $E_T = 0.001$. The parameters of HMM(4) are given explicitly in table 9.

Denote by $F^0_{ij}(x)$ the cumulative distribution function

$$F^0_{ij} = \begin{cases} \Pr[0 < f_i(O_T) < x \mid O_T \in \text{HMM}(j)] , & \text{for } x > 0 , \\ 0, & \text{for } x \leq 0 . \end{cases} \tag{33}$$

Ten-thousand observation sequences O_T were generated from HMM(4) for T = 25. As given in table 8, 121 sequences resulted in zero posterior likelihood function values (that is, $f_3(O_T) = 0$) and the remaining 9879 nonzero values of $f_3(O_T)$ give the histogram shown in figure 6. By comparison with figure 4, it is clear that no significant difference between log $dF^0_{34}(x)$ and log $dF_{33}(x)$ is evident. Therefore, the misclassifi-

Table 9. Parameters of HMM(4), Rounded to Three Significant Digits

NUMBER OF MARKOV STATES = 5

NUMBER OF SYMBOLS PER STATE = 8

INITIAL STATE PROBABILITY VECTOR:

1.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00

TRANSITION PROBABILITY MATRIX:

6.00E-01	4.00E-01	0.00E+00	0.00E+00	0.00E+00
0.00E+00	7.00E-01	2.00E-01	1.00E-01	0.00E+00
0.00E+00	0.00E+00	6.00E-01	4.00E-01	0.00E+00
0.00E+00	0.00E+00	0.00E+00	7.00E-01	3.00E-01
0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.00E+00

SYMBOL PROBABILITY MATRIX (TRANSPOSED):

8.99E-01	1.00E-01	1.43E-04	1.43E-04	1.43E-04
1.00E-01	5.99E-01	1.43E-04	1.43E-04	1.43E-04
1.43E-04	2.00E-01	3.00E-01	1.43E-04	1.43E-04
1.43E-04	1.00E-01	5.99E-01	1.00E-01	1.43E-04
1.43E-04	1.43E-04	1.00E-01	2.00E-01	1.43E-04
1.43E-04	1.43E-04	1.43E-04	4.00E-01	1.00E-01
1.43E-04	1.43E-04	1.43E-04	3.00E-01	5.99E-01
1.43E-04	1.43E-04	1.43E-04	1.43E-04	3.00E-01

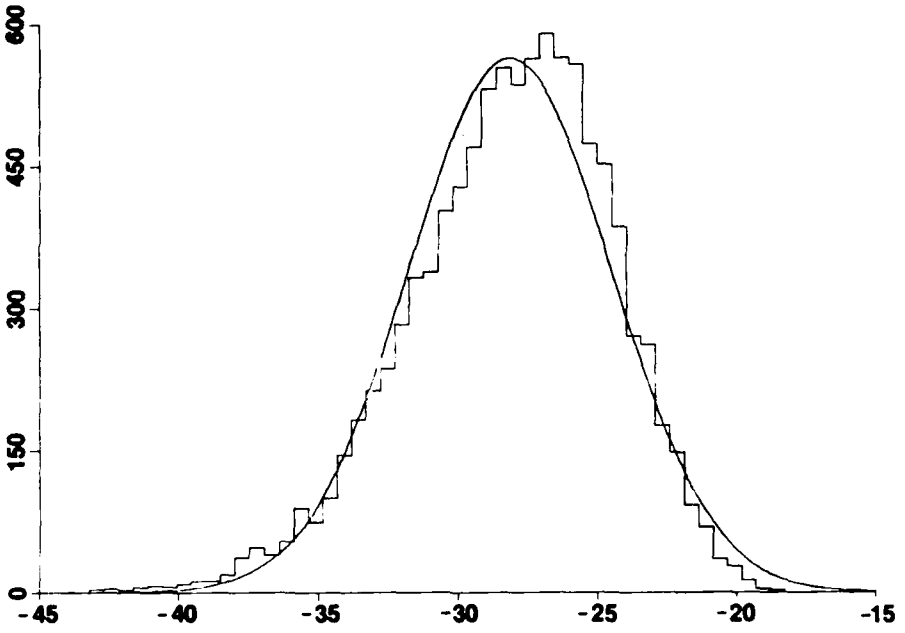


Figure 6. Histogram of 9879 Samples of log $dF_{34}^{0}(x)$ for T = 25. (The Normal Curve Has the Sample Mean = -28.156 and the Variance = 3.6167.)

cation rate due to noise-induced shifts in the statistics of $dF_{34}^{0}(x)$ is effectively zero. The maximum posterior classifier is thus 98.8 percent reliable when used with noisy observations characterized by the noise probability matrix H.

Because $E_T = 0.001$ in this example, each observation sequence O_{25} has probability 0.025 of having at least one incorrect symbol. Of 10000 observation sequences, the expected number with at least one incorrect symbol is 250. Nearly half (121) contained forbidden symbol sequences and caused the only significant misclassification problem. The other half apparently made no contribution to misclassification.

It would be desirable to be able to compute the moments of $F_{ij}^{0}(x)$ instead of $F_{ij}(x)$. Alternatively, it would be desirable to be able to compute the amplitude of the impulse (delta function) in $dF_{ij}(x)$ that seems to be present in the left-to-right HMMs considered here. In other words, if we write

$$dF_{ij}(x) = A\,\delta(x) + dF_{ij}^{0}(x) \quad , \tag{34}$$

then an algorithm to compute A directly would be worthwhile. Knowing A and the moments of F_{ij} gives the moments of $F^{0}_{ij}(x)$. However, developing such an algorithm requires further work.

IV. CONCLUSIONS

The first two moments $M_{ij}(1,T)$ and $M_{ij}(2,T)$ of $F_{ij}(x)$ give good estimates of the probability density function $dF_{ij}(x)$ in the case when HMM(i) and HMM(j) are both ergodic. The reason is that $dF_{ij}(x)$ is apparently nearly log-normal. The evidence supporting this claim is strictly empirical and a proof of the degree of approximation of $dF_{ij}(x)$ to log-normal would be worthwhile. When either HMM(i) or HMM(j), or both, are not ergodic, $dF_{ij}(x)$ does not approximate the log-normal distribution. Consequently, higher order moments are needed to develop reasonable continuous approximations to $dF_{ij}(x)$.

REFERENCES

1. L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent Isolated Word Recognition," The Bell System Technical Journal, vol. 62, no. 4, April 1983, pp. 1075-1105.

2. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," The Bell System Technical Journal, vol. 62, no. 4, April 1983, pp. 1035-1074.

3. H. L. Van Trees, Detection, Estimation, and Modulation Theory, Part I, chapter 2, Wiley and Sons, New York, 1968.

4. A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, 1965, p. 158.

5. L. A. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," IEEE Transactions on Information Theory, vol. IT-28, no. 5, September 1982, pp. 729-734.

6. J. G. Kemeny and J. L. Snell, Finite Markov Chains, chapter 3, Van Nostrand, Princeton, New Jersey, 1960.

7. C. E. Shannon, "A Mathematical Theory of Communication," The Bell System Technical Journal, vol. 27, 1948.

INITIAL DISTRIBUTION LIST

Addressee	No. of Copies
NAVSEA (Code 63-D (CDR E. Graham, C. Walker))	2
NRL (Code 5130 (Dr. P. Abraham), - 5844 (Dr. F. Rosenthal, Dr. R. J. Hansen), -5104 (Dr. S. Hanish), -5130 (R. Menton, J. Cole))	6
NRL/USRD, Orlando (A. Markowitz)	1
NOSC, San Diego (E. Schiller, R. Johnson, R. Smith, G. Benthein)	4
DTNSRDC, Bethesda (Code 1541 (P. Rispin))	1
DARPA (CMDR A. Sears)	1
DTIC	12
NAVPGSCOL, Monterey	1
Naval Technology Office (L. Epley)	1
ONR (Code 434 (Dr. N. Glassman, Dr. R. Fitzgerald), -220B (Dr. T. Warfield))	3
ONR, Boston (Dr. R. L. Sternberg)	1
AMSI (W. Zimmerman)	1
Bell Laboratories (S. Francis, G. Zipfel)	2
S. Berlin, Inc. (S. Berlin)	1
Campbell and Kronauer Associates (C. Campbell, R. Kronauer)	2
Gould, Inc. (S. Lemon)	1
G&R Associates (F. Reiss)	1
SCRIPPS (D. Andrews, Jr.)	1
Technology Service Corp. (L. Brooks)	1
Atlantic Applied Research (S. Africk)	1
Signal Technology, Inc. (Dr. S. B. Davis)	1
ATT Bell Laboratory (Dr. S. Levinson)	1
Digital Equipment Corp. (Dr. T. Vitale)	1
Atlantic Aerospace Electronics Corp. (L. Lynn)	1
URI, Dept. of Math. (Prof. P. T. Liu)	1
URI, Dept. of Elec. Eng. (Dr. D. Tufts)	1
Chase, Inc. (D. Chase)	1
A&T, No. Stonington (L. Gorham)	1